ABSTRACT

Methods and systems for information extraction are disclosed. In one such method and system, a sample of related articles is obtained, and an article is selected as a seed article. The distances between sample articles are calculated to determine a set of one or more closest articles to the seed article. The set of closest articles is used to identify information fields containing variable data within the seed article. There are a variety of techniques by which this may be performed, one of which is by using dynamic programming alignment to compute alignments between articles. The information fields are labeled, and a template is generated using the labeled fields. The template is used to extract data from a source article by comparing the source article with the template and associating the variable data of the source article with the labeled fields.